Comparative and functional analysis of cardiovascular-related genes

Jan-Fang Cheng[†1,2] *&*
Len A Pennacchio[1,2]

[†]*Author for correspondence*
[1]*Department of Genome Sciences, MS 84-171, One Cyclotron Road, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
Tel: +1 510 486 6590;
Fax: +1 510 486 6635;
E-mail: jfcheng@lbl.gov
[2]*US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA*

The ability to detect putative *cis*-regulatory elements in cardiovascular-related genes has been accelerated by the availability of genomic sequence data from numerous vertebrate species and the recent development of comparative genomic tools. This improvement is anticipated to lead to a better understanding of the complex regulatory architecture of cardiovascular (CV) genes and how genetic variants in these non-coding regions can potentially play a role in cardiovascular disease. This manuscript reviews a recently established database dedicated to the comparative sequence analysis of 250 human CV genes of known importance, 37 of which currently contain sequence comparison data for organisms beyond those of human, mouse and rat. These data have provided a glimpse into the variety of possible insights from deep vertebrate sequence comparisons and the identification of putative gene regulatory elements.

Introduction

Cardiovascular diseases (CVD) are frequently the result of environmental factors (e.g., diet, smoking, sedentary lifestyle) coupled with less well-defined genetic contributors that together determine an individual's CVD susceptibility. It is estimated that half of all American deaths are the result of CVD, which includes a complex set of disorders such as coronary heart disease, stroke, hypertensive disease and rheumatic heart disease. Cumulatively, CVD represents a huge medical burden with an estimated cost of US$350 billion in the year 2003 alone [101]. In the past several decades, a large number of studies have focused on the identification and understanding of genes that contribute to the progression of CVD. While significant progress has been made in identifying some of the key protein-encoding participants, little is known about the mechanisms that determine the level, location and chronology of expression of potentially important susceptibility genes. One key reason for this lack of information is an inability to computationally identify gene regulatory sequences embedded in the > 95% of the human genome that does not code for proteins. This is in stark contrast to gene-predictions where significant progress has been made in the last 5 years.

With the availability of the finished human genome sequence [102] and several other vertebrate genome drafts [1,2,103,104], comparative genomics has revealed a large number of highly conserved non-coding sequences with putative functionality. This assumption is based on the simple hypothesis that conserved sequences are functionally important due to evolutionary constraints that have selected against mutations within these sequences. The conserved DNA pieces include function such as protein encoding exons, non-coding RNA genes, chromosome structural elements and gene regulatory sequences [3].

In this review, we focus on a single database [105] that encompasses multiple cross-species sequence comparison data for a set of human CV related genes. This resource is publicly available and provides an entry point for additional genomic information for CVD genes and their possible *cis*-regulatory sequences. Rather than providing a summary of all publicly available comparative genomic databases and resources, this review aims to provide early insights of what has been learned from such data sets. More details on additional comparative genomic tools and databases can be found in other recent reviews [4-6].

CV genes and comparative sequence analysis

As part of a 'Programmes for Genomics Applications' (PGA) sponsored by the US National Heart, Lung and Blood Institute, the Berkeley PGA has sought to provide insights into possible gene regulatory elements in the vicinity of CV genes. This resource is focused on providing the scientific community with comparative sequence analysis from a range of vertebrates for 250 well-studied CV genes with the goal of adding highly

Keywords: apolipoprotein, BAC, cardiovascular, comparative genomics, gene regulation, orthology, PGA, phylogenetic shadowing, VISTA

Ashley Publications Ltd
www.ashley-pub.com

conserved non-coding elements as another layer of possible functional elements in the proximity of these genes. These genes were selected using the following criteria:

- disease genes – genes in which null or structural mutations have been shown to contribute to CVD
- gene clusters – clustered genes that arose as the result of ancient gene duplication events and may share regulatory elements
- genes or regions with interesting regulatory features – genes exhibiting differences in expression within a species and between species (e.g., *APOA1*, *CYP7A1*, and *APO(a)*)
- input from experts in the cardiovascular field

The identified conserved elements could subsequently serve as a data set for *in vivo* studies to determine if these elements possess *cis*-regulatory activity. The genes selected for this analysis are categorized in six general areas that comprise:

- heart and vascular development
- blood pressure/kidney
- blood pressure/homeostasis
- atherosclerosis (vascular biology/lipoprotein metabolism)
- thrombosis
- hypertrophy/heart failure

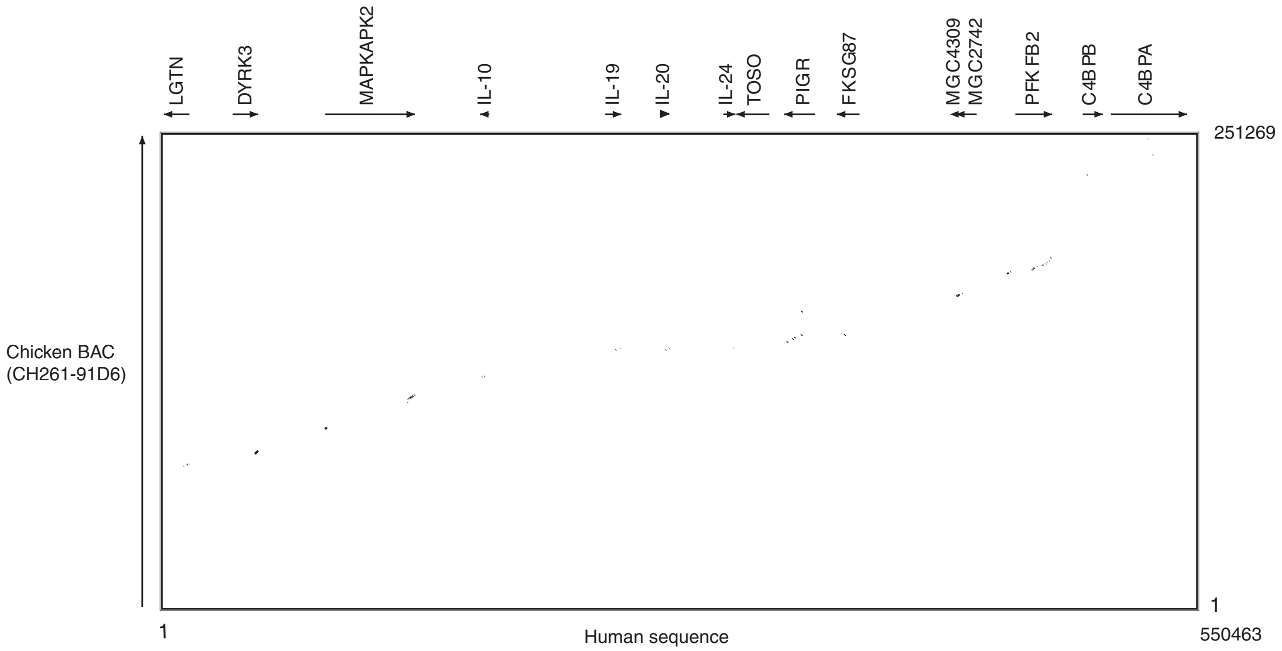
Isolation of bacterial artificial chromosomes and verification of orthologous sequences

One of the first challenges in comparative sequence analysis is the determination of which species to compare. In general, human–mouse comparisons have been used as the benchmark to assess species for subsequent DNA sequencing. While several vertebrate whole genome sequences are beginning to become available (human, mouse, rat, pufferfish, zebrafish etc.), genome sequences of many other vertebrate species, despite their utility in comparative sequence analysis, are limited. Clone-based target sequencing is often necessary to obtain complete sequence data for a given gene of interest. Consequently, large insert clones of bacterial artificial chromosomes (BACs) have become the standard substrates for targeted DNA sequencing [7,8]. Most of the publicly available BAC libraries for acquisition can be found on the web [106,107]. In general, BAC libraries are arrayed onto membrane filters that are screened by hybridization with short 'overgo probes' [9] derived from cDNA sequences of the targeted species. It is important to avoid sequences that are shared by pseudo-genes or other members of the same gene family. For species that do not have cDNA sequences available, probes from other related species may be used in the hope that highly conserved DNA segments will cross hybridize with the gene of interest.

Once positive clones are identified, it is important to verify their orthology (i.e., DNA in different species are descended from the same piece of DNA in the last common ancestral species) since orthologous genes usually preserve similar functions and gene regulation that are reflected in the sequence conservation. On the other hand, non-orthologous sequences do not preserve these features. One way to verify orthology is to see if the neighboring genes are also present in the isolated BACs because orthologous sequences have evolved as genomic blocks that usually contain multiple genes. Therefore, verification of orthologous sequences within BACs can be conducted by polymerase chain reaction (PCR) or hybridizations using probes derived from neighboring genes.

The comparison of conserved genomic segments of 40 mammalian species at the chromosomal level [10] and the comparison of human and mouse genomes at the sequence level [1] have clearly demonstrated the preservation of long-range gene order across mammalian species. This type of genomic segmental conservation is also seen in human and chicken comparisons but is significantly smaller in human–fish comparisons [2]. In our data set, more than 10 genomic intervals with more than one gene in the chicken BAC clone have shown the same gene order and orientation as in humans, which agrees with previous observations [11]. One such example is shown in **Figure 1**, where a 251 kb chicken BAC containing 15 genes is in perfect synteny with the orthologous human genes in both order and orientation. Occasionally, we have observed evolutionary breakpoints of chromosomes within a BAC. In those cases, additional species sequence comparisons will be useful for determining the ancestral gene order prior to the rearrangement(s). Comparison of the human genome to that of species with more remote ancestry such as the pufferfish, *Fugu rubripes* [2], has revealed a much shorter degree of segment conservation that hinders this criterion to verify orthology. In the case of human and fish, reciprocal best matches of both DNA and protein sequences have been used as an alternative way of verifying orthology. Sequence alignment tools and servers useful in

Figure 1. A conventional dot plot of two orthologous sequences.



The x-axis represents the human sequence spanning from nucleotides 203,415,554 to 203,966,016 of chromosome 1 based on the NCBI Build 33. The y-axis represents draft sequence obtained from a chicken BAC that was isolated using an *IL-10* 'expressed sequence tag' sequence probe. Position and orientation of the genes on the top of the plot are based on the human sequence. The dot plot was created by PIPMAKER [110] and the dots indicate regions of sequence homology.

determining orthologous genes or aligning sequenced intervals can be found in **Table 1**.

The VISTA comparative genomic tool

Many sequence comparison tools have been developed and improved in the past few years [4-6] to meet the demand of analyzing progressively increasing amounts of genomic sequence data. As a part of the Berkeley PGA, we have participated in the development of the VISualization Tool for Alignment (VISTA) software and have utilized this method to display the CV gene comparative sequence data set. Other suitable tools include PipMaker [12], SynPlot [13], Alfresco [14], and GLASS [15]. While this is not intended to be a detailed overview of the VISTA program, a few important features of the VISTA tool for scoring conserved sequences are highlighted. First, the VISTA plot uses sequence alignment outputs generated by AVID, a global alignment program [16]. VISTA has recently utilized a newly developed global alignment program called LAGAN [17], which was also used in the human versus rat genome comparison [108]. This is in contrast to several other programs, which use local alignment algorithms such as BLAST (Basic Local Alignment Search Tool). Global alignment programs assume that the two sequences being aligned have all their functional elements preserved in the same order and orientation and calculate a best match across the entire length of the alignment. The outcome of this assumption is better specificity at detecting conserved sequences with true biological significance, but, at the same time, the program may miss regions where inversion/rearrangements have occurred. A new alignment tool, shuffle-LAGAN, has been developed to address this issue [18]. Second, the VISTA software uses a sliding window to calculate percent identity over a specified window length at each base pair and then draws a continuous curve to display levels of identity [19]. Such a plot-like graphical display captures conservation in non-coding regions more effectively than other available dot plot style displays since gaps are averaged over the length of the alignment. However, the curve display does not readily distinguish genic and non-genic conservation – dot plots are better able to differentiate this based on the perfect co-linear relationship of open reading frames. Finally, the VISTA software provides options to combine

Table 1. Computational tools and servers useful in determining orthologous genomic intervals and generating sequence alignments.

Tools	Websites
To search for the best sequence matches in the completed and drafted genomes	
BLAST (NCBI)	http://www.ncbi.nlm.nih.gov/BLAST/
GenomeVISTA	http://pipeline.lbl.gov/cgi-bin/GenomeVista
BLAT (UCSC Genome Browser)	http://genome.uscs.edu/cgi-bin/hgBlat
SSAHA (EMSEMBL)	http://www.ensembl.org/
To align two or more segments of genomic sequences	
BLAST2 Sequences	http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html
AVID	http://www-gsd.lbl.gov/VISTA/details_avid.htm
LAGAN	http://lagan.stanford.edu/
PIPMAKER	http://bio.cse.psu.edu/pipmaker/
To download human sequence annotation	
UCSC Genome Browser	http://genome.uscs.edu/

clustering of predicted transcription factor binding sites (TFBSs) and the analysis of inter-species sequence conservation to maximize the identification of functional sites [20].

With the above criteria of determining orthologous genes, we have generated multiple species sequence data from organisms ranging from fish to chimpanzee in 25 genomic intervals containing 37 CV-related genes (Table 2). Of these, 30 genes have been shown to play an important role in cholesterol transport and metabolism. The VISTA sequence comparison of these genes can be found at [105] using a 'Reference Sequence' (RefSeq) gene name search. This data set is beginning to be used to identify putative *cis*-regulatory elements based on conservation and some of the insights derived from this preliminary analysis are discussed in the following sections.

No two species reveal all evolutionarily conserved sequences

Sequence comparisons between human and mouse, which diverged ~ 80 million years ago, have revealed conservation in most coding exons and also in a large number of non-coding sequences [1,21-22]. Despite the early success of identifying functional non-coding elements through human–mouse comparisons [23-25], the comparison of two distantly related species is clearly not sufficient to detect species-specific functional elements, such as genes, exons and regulatory sequences, that only recently emerged since their last common ancestor. In contrast, comparing two closely-related species (such as primates) to uncover shared conserved intervals poses a problem due to the large amount of background conservation that is indistinguishable from functional conservation. Another factor in deciding which species to compare is the evolutionarily distance as calculated by the rate of nucleotide substitution between two species. There is significant evidence showing that mammalian genomes accumulate silent mutations at different rates across different regions of the genome [26]. Although not intuitive, the mutation rate of any particular interval may affect the outcome of the alignment as much as the species selected for comparison. We will use two examples of CV gene sequence comparisons to demonstrate how adding a new species to the comparison aids in ameliorating the problem of 'too much' or 'too little' conservation. The first example is the low-density lipoprotein receptor gene (*LDLR*) that encodes a cell surface receptor that plays an important role in cholesterol homeostasis. Human–mouse comparison of *LDLR* suggests a fast-evolving genomic interval where only one 142 bp conserved non-genic element has significant similarity in a 40 kb region spanning ~ 17 kb of the $5'$ flanking region and the first 8 exons of the gene (Figure 2A). In this comparison, the criterion for detecting sequence similarity was arbitrarily set at 75% identity over 100 bp. This single human–mouse conserved sequence, located immediately upstream (-4 to -145 bp) of the *LDLR* transcription start site, contains a well characterized sterol regulatory element-1 (SRE-1) that is responsible for transcription activation of *LDLR* by conditional positive enhancer proteins, SRE-binding proteins (SREBPs) [27]. However, another known *cis*-regulatory element upstream of *LDLR* (located at -255 to -139 bp), which luteinizing hormone (LH) and insulin/IGF-1 act upon, is not found within a conserved interval [28]. Thus, human–mouse comparison for *LDLR* is sufficient to detect some *cis*-regulatory elements based on conservation but fails to detect others even with a reduced stringency of 65% identity over 100 bp. In an attempt to look for other possible conserved non-coding sequences that may explain other aspects of *LDLR*'s expression pattern, comparison of human with lemur (a prosimian that diverged ~ 55 Myr) [29] was performed. As shown in Figure 2A, numerous additional non-

Table 2. Multiple species sequence comparison for the CV genes at the Berkeley PGA. These species represent those in addition to human–mouse

Gene targets	Organisms	GenBank	GenBank
ABCA1	Lemur	AC139880	AC140021
ABCG1	Lemur	AC145532	
	Rabbit	AC145541	
ABCG5/G8	Lemur	AC145533	
	Rabbit	AC093410	
	Fugu	AC146282	
	Titi	AC146286	
ACE	Lemur	AC118572	
	Rabbit	AC118582	
	Chicken	AC118566	AC145528
	Opossum	AC118578	
APOA1/C3/A4/A5	Chimp	AC113242	
	Baboon	AC145521	
	Marmoset	AC145529	
	Titi	AC144989	
	Lemur	AC118574	
	Rabbit	AC118580	
	Opossum	AC145538	
	Chicken	AC110875	
	Xenopus	AC146287	
	Zebrafish	AC146294	
APOB	Baboon	AC140975	
	Lemur	AC118571	
	Rabbit	AC118581	
	Opossum	AC145539	
	Chicken	AC120501	
APOC2/C4/C1/E	Chimp	AC120211	
	Baboon	AC145523	
	Marmoset	AC146283	
	Titi	AC146285	
	Lemur	AC135911	
	Zebrafish	AC146288	
APOL1/L2/L4/L3	Baboon	AC145524	AC145525
	Lemur	AC145535	
APOL5/L6	Baboon	AC145526	
	Lemur	AC145534	AC145536
CETP	Baboon	AC120499	
	Marmoset	AC145462	
	Titi	AC142393	
	Lemur	AC140976	
	Rabbit	AC145464	
CNN1 (calponin)	Rabbit	AC145542	
	Opossum	AC144892	
CYP7A	Rabbit	AC110419	
F2RL3	Lemur	AC118570	
GATA4	Opossum	AC144669	AC145537
	Chicken	AC110874	
IL10	Opossum	AC145194	
	Chicken	AC145193	
LCAT	Baboon	AC145522	

Table 2. Multiple species sequence comparison for the CV genes at the Berkeley PGA. These species represent those in addition to human–mouse (continued)

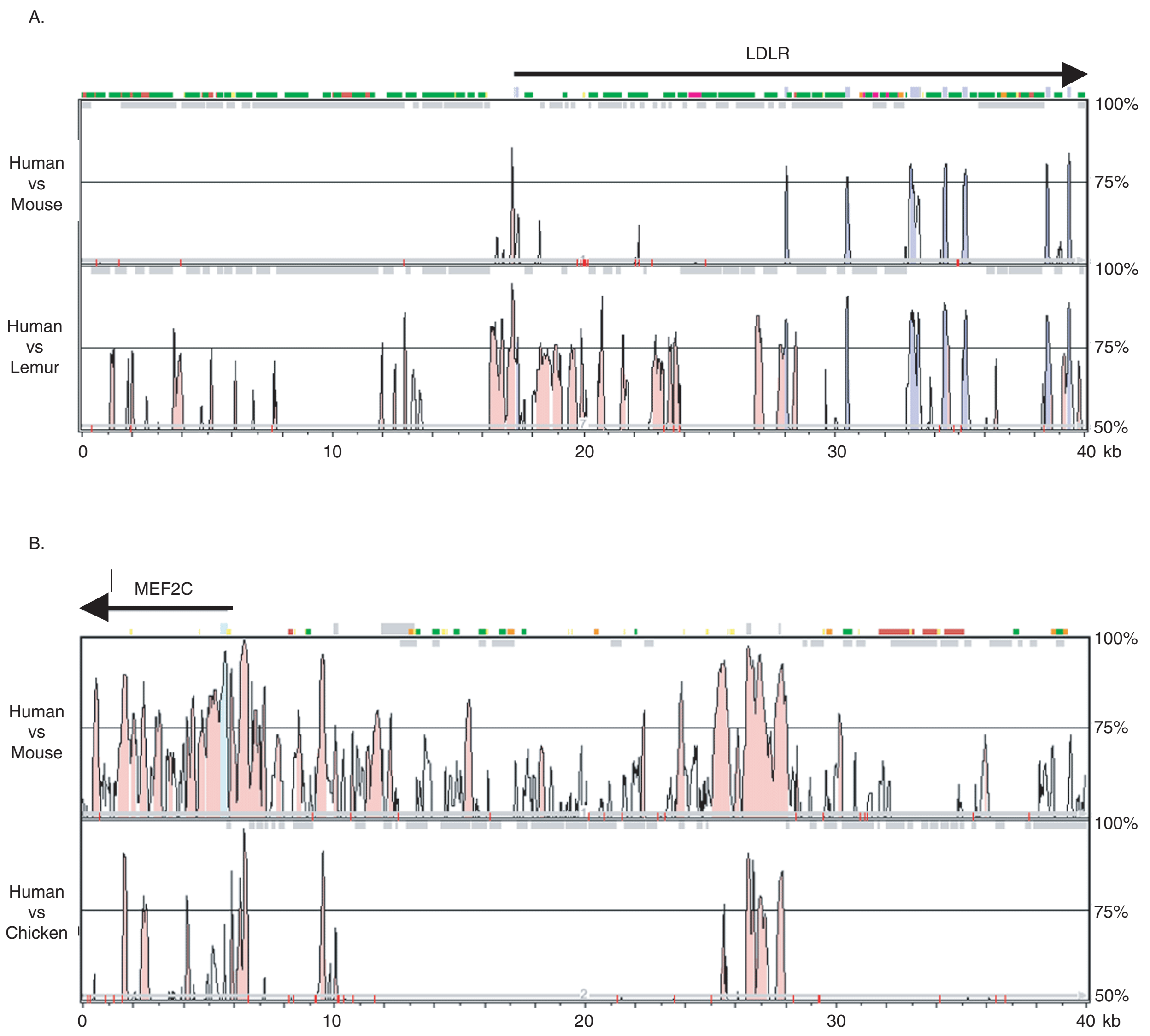
Gene targets	Organisms	GenBank	GenBank
LDLR	Baboon Marmoset Titi Lemur	AC140974 AC145530 AC144655 AC118569	
LXRA	Baboon Lemur Rabbit Hedgehog Opossum Chicken	AC140973 AC118575 AC110879 AC122113 AC118577 AC118567	
MEF2C	Opossum Chicken Zebrafish	AC145288 AC118565 AC146289	 AC145020
PLG/APO(a)	Chimp Lemur Hedgehog	AC132187 AC093405 AC122114	 AC131892
PPARA	Lemur Hedgehog Opossum Chicken	AC118573 AC131890 AC118579 AC118568	
PPARG	Lemur Rabbit Hedgehog Opossum Chicken	AC145531 AC131898 AC142242 AC131896 AC131893	
SCD	Baboon Lemur	AC139668 AC139669	
SREBF1	Lemur Hedgehog Opossum Chicken Rabbit	AC141085 AC145527 AC131895 AC144804 AC145540	
SREBF2	Lemur Rabbit	AC146284 AC145540	

CV: Cardiovascular; PGA: Programmes for Genomics Applications.

coding conserved sequences were found that fit the 75% identity over 100 bp criterion. These conserved sequences include the previously known -255 to -139 bp *cis*-regulatory element, which was missed by human–mouse sequence comparison alone. This demonstrates the value of performing sequence comparisons of evolutionarily 'closer' species when human–mouse comparisons fail to detect both known and candidate non-coding conserved elements.

A second example where human–mouse sequence comparison was not well-suited to reveal gene regulatory elements is the *MEF2C* gene, a member of the MADS box transcription enhancer gene family, that encodes a helix-loop-helix protein involved in myocyte differentiation [30]. In contrast to *LDLR*, the genomic interval of *MEF2C* represents a slow-evolving region where the human and mouse comparison reveals 35 conserved non-coding elements in a 40 kb region spanning ~ 34 kb of the 5′ flanking sequence and the first exon of the gene (Figure 2B). In an attempt to reduce the amount of non-coding conservation prior to biological experimentation, a more distant sequence comparison to human was performed. Comparison of *MEF2C* in human and chicken provided significantly fewer conserved non-coding sequences, yielding a more manageable number of candidate gene regulatory elements for *in vivo*

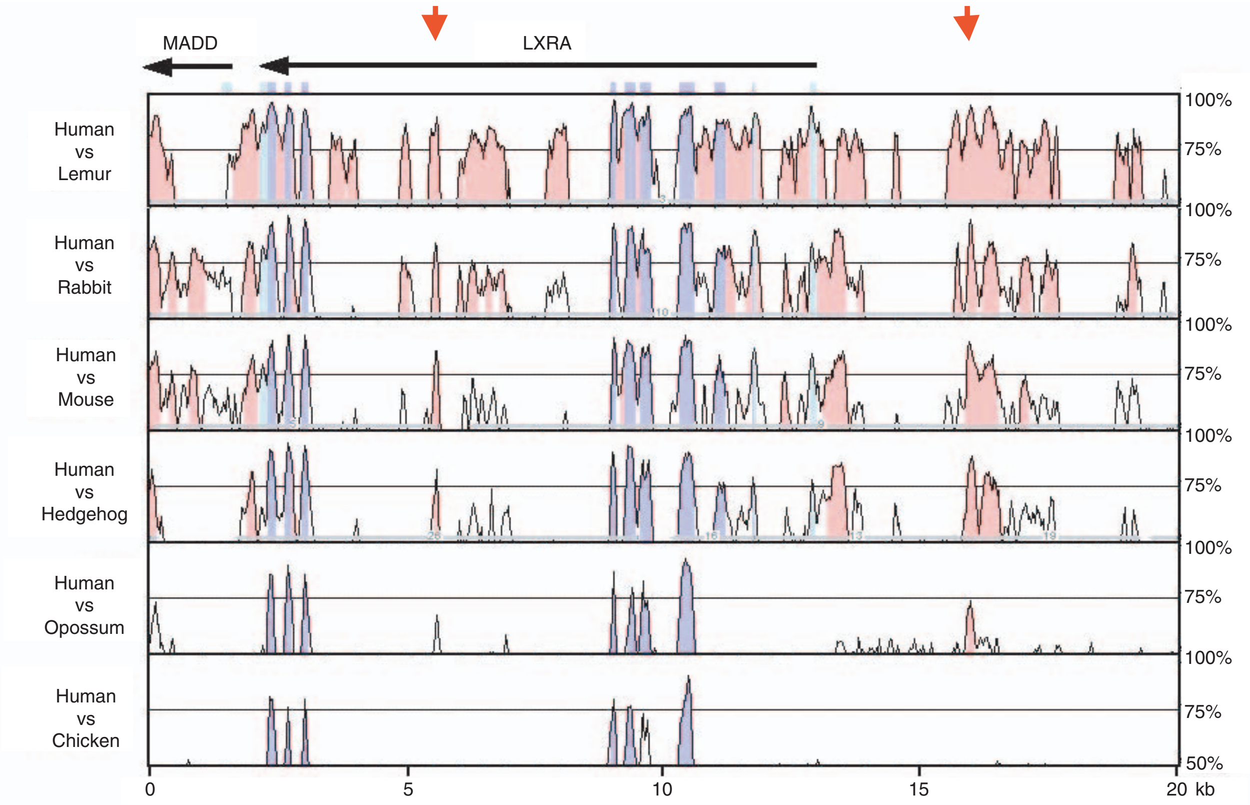
Figure 2. Two examples of VISTA plots showing regions with low- or high- sequence homology between human and mouse.



Two human sequences indicated by the horizontal axis, one containing the 5′ portion of the *LDLR* gene (**A**) and the other containing the 5′ portion of the *MEF2C* gene (**B**). For the *LDLR* gene, human–lemur comparison is also provided (A, lower panel). For the *MEF2C* gene, human–chicken comparison is displayed (B, lower panel) Horizontal arrows indicate the direction of transcription for each gene. The locations of coding exons and UTRs are shown as blue and turquoise rectangles above the profile, respectively. The smaller color rectangles above the profile show the locations of repetitive sequences. Conserved sequences are represented by peaks where blue, turquoise, and pink color peaks represent coding, UTRs and non-genic sequences, respectively. Pair-wise percent identities are indicated on the vertical axis.

studies. An alternative way of prioritizing conserved elements would be to select the few most highly conserved human–mouse non-coding sequences, but this strategy, unlike multiple cross-species comparison, does not add confidence to distinguish selective from background conservation. Indeed, recently one of the human–chicken *MEF2C* conserved non-coding elements has been experimentally confirmed to drive skeletal muscle expression in transgenic mice and a second human–chicken conserved element has led to the identification of a cardiac muscle enhancer, consistent with *MEF2C's* known expression pattern (Brian Black, UCSF,

Figure 3. Multiple pair-wise comparison of the *LXR*-α gene and flanking sequences demonstrating peaks of common conservation in a multiple VISTA plot.



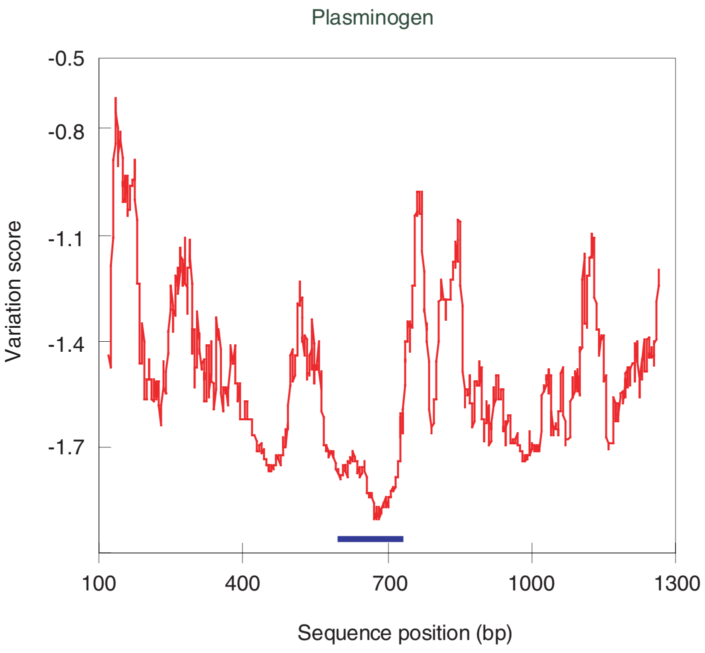
The x-axis represents the human sequence that is being compared with six other vertebrates. The definition of arrows, rectangles, and peaks are the same as that in the legend for **Figure 2**. The vertical arrows indicate the location of two non-coding elements conserved throughout mammals.

personal communication). Thus, comparisons of more distantly related species can also be useful for prioritizing conserved elements for gene regulatory functional studies. These examples highlight the importance of selecting the species for comparison based upon the genomic interval under investigation.

The advantage of multiple cross-species comparisons

In the above examples, human–mouse comparative data were examined and leveraged to make hypotheses about which species to sequence next to maximize sequence information. For comparisons where 'too much' conservation was detected, more distant species were sequenced, while for comparisons where 'too little' conservation was detected, more closely related species were compared. This trial and error strategy has proven successful but can take several iterations to find the right window of species for informative comparisons. As an alternate approach, a wide range of vertebrate species can be sequenced in parallel without making prior assumptions based on human–mouse comparative data. With the increasing availability of BAC libraries covering a wide range of phylogenetic distances, multiple cross-species sequence comparisons of a large segment of genomic DNA is no longer a difficult task. One example where this was applied is in the analysis of the *LXR*-α gene (**Figure 3**). For this gene, six vertebrate sequences were compared with human in a 20 kb interval containing the entire *LXR*-α gene and 7 kb of 5′ and 2 kb of 3′ flanking sequence. In macrophages, *LXR*-α stimulates the transcription of genes encoding transporters involved in cholesterol efflux, which may prevent the transformation of these cells into foam cells in response to lipid loading [31]. Previous analysis of the *LXR*-α promoter using

Figure 4. Primate-specific 'phylogenetic shadowing' reveals a previously defined exon of the plasminogen gene.



Sequences derived from eight primates were compared with the human sequence. On the y-axis, a variation score is provided with more negative scores indicating less variable regions, and on the x-axis 1300 bp of human sequence is displayed. The method of calculating negative scores has been previously described [34]. The known plasminogen exon in this interval is depicted by a solid black line within the plot. Note the decreased amount of primate variation in regions corresponding to the exon with known functional importance. The figure was kindly provided by Dr Boffelli, Lawrence Berkeley National Laboratory.

reporter constructs in HepG2 cells and gel-shift analysis revealed a functional LXR/RXR binding site ~ 2.9 kb upstream of the transcription initiation site [32]. This *cis*-regulatory element is believed to be involved in the amplification of the effects of oxysterols on reverse cholesterol transport and is therefore a potential therapeutic target for the treatment of atherosclerosis. The multiple vertebrate cross-species comparison revealed at least two non-coding elements conserved throughout mammals, including one located at 2.9 kb upstream of the gene (**Figure 3**). Thus, this highly conserved element strongly overlaps with a previously defined *LXR*-α gene regulatory element and further supports such a strategy to identify novel gene regulatory elements.

Another advantage of multiple species comparison is the more precise definition of possible functional elements within these conserved sequences. The *LXR*-α -2.9 kb element is ~ 1000 bp in size when examined in the human–mouse comparison but this is significantly reduced to ~ 200 bp in the human–opossum comparison. Thus, these multiple species sequence alignments suggest that this 200 bp conserved sequence present in all mammals contains critical *cis*-regulatory elements, providing a method to scan for possible TFBSs within a fairly small window. A third advantage of multiple-species comparisons is the evolutionary insights from the various conserved sequences throughout the targeted region. For example, the 5′ coding exons in the *LXR*-α gene are less conserved compared to the 3′ coding exons as seen in the human–opossum and human–chicken comparison, supporting a faster rate of mutation accumulation in the amino-terminus of the protein. Indeed, protein comparisons reveal the ligand binding domain in the carboxy-terminus of *LXR*-α (aa 215 to 434, SwissProt: locus NRH3_HUMAN, accession Q13133) is highly conserved relative to the amino-terminus, further supporting the functional constraints on this region of the protein. This type of sequence analysis may be used to speculate on conserved functional domains in protein-encoding regions with unknown function.

Expert opinion and outlook

We are just beginning to discover the realm of possible insights derived from comparative data, and an exponential increase in vertebrate genomic sequence is expected in the not so distant future. Strategies to deal with these large data sets represent a challenge and significant efforts focused on exploiting this information are expected to be a fruitful area of study. In the concluding sections, two broad areas likely to increase in use in the comparative genomic field will be described. It is anticipated that improvement in these areas will enhance our ability to recognize the functional information embedded in the human genome.

Multiple primates provide sufficient diversity to detect functional conservation

Comparison of distantly related genomes (such as human–mouse) are ill-equipped to reveal functional elements limited to more closely related species (such as primates). For instance, alternatively spliced exons in the human, mouse and rat genomes are poorly conserved relative to constitutively spliced exons [33]. It is estimated that 40–60% of the human genes consist of alternatively spliced variant, and 72% of the

Highlights

- Complete or nearly complete genomes of five vertebrates (human, mouse, rat, pufferfish, and zebrafish) are beginning to provide insights into human biology.
- Several comparative sequence analysis tools have been developed to align whole genomes and enable the identification of conserved coding and non-coding sequences. These data provide the substrates for both custom- and large-scale functional studies.
- Multiple species comparison increases the confidence of predicting functionally conserved sequences that have undergone selective constraint.
- The biological questions being addressed by comparative analysis usually dictate the appropriate species for analysis. Therefore, having the ability to generate targeted genomic sequences from various vertebrates is important.
- 'Phylogenetic shadowing' has demonstrated that sequences derived from 5–7 primate species provide sufficient diversity to detect functional conservation. These data support the theory that multiple primate sequence comparisons may be a major method of detecting functional elements specific to primates that may be relevant to human disease.

alternatively spliced exons are present only in humans or in mice and therefore will not be detected by human–mouse sequence comparisons. Thus, it is reasonable to postulate that a significant fraction of human regulatory sequences are also primate-specific and will not be detected by human–mouse comparisons. This observation illustrates the need for strategies to study closely-related species that share common biological traits. In a recent sequence comparison study using primates, Boffelli *et al.* (2003) illustrated that cumulative sequence differences from a collective number (5–15 species) of New World monkeys (40 Myr), Old World monkeys (25 Myr), and Apes (6–14 Myr) can reveal functionally conserved elements corresponding to exons and putative regulatory sequences (**Figure 4**). The principle behind this strategy, termed 'phylogenetic shadowing', is to compare orthologous sequence from numerous primate species to increase the total evolutionary distance being examined. Instead of the traditional pairwise comparison (such as human–mouse), 'phylogenetic shadowing' compares a dozen or more different primate species. The additivity of these primate differences robustly defines regions of increased variation and 'shadows' representing conserved segments. While this approach was tested on only four exons and one promoter region, it is expected to be particularly useful for the sequence analysis of certain CV genes such as *CETP*, *Apo(a)*, *ApoL1-4*, and *LDLR* genes where little or no sequence conservation is found in human–mouse comparisons. With efforts underway to sequence the chimpanzee and macaque genomes (and likely more primates on the way), this type of strategy is expected to contribute to our understanding of primate-specific functional elements in the human genome.

Comparative genomics and the integration of other genomic datasets

Mapping of evolutionarily conserved sequences provides a new landmark of potentially functional elements in the human genome. Parallel efforts have provided the identity of transcribed portions of the genome ('expressed sequence tags') and the location of human single nucleotide polymorphisms (SNPs). These complementary data sets are beginning to be integrated and are expected to provide an added layer of functional information. One example of where these data sets have been merged is the UCSC Genome Browser [109]. While this is an immature field, these resources provide the framework for a better annotation of the functional portion of human genome.

In this review, we have described the availability and importance of a database containing multiple-species sequence comparisons for a set of CV genes. We have used several gene-by-gene examples to demonstrate the power of vertebrate sequence comparisons to reveal biologically functional sequences. This particular field is just beginning to gain momentum and should provide important insights into gene regulatory sequences embedded in the large non-coding portion of the human genome. Such information may provide clues to gene regulatory defects in CVDs and potential therapeutic entry points through the modulation of CV gene expression.

Acknowledgments

We thank Dario Boffelli, Marcelo Nobrega, and James Priest for critical reading of the manuscript and the Berkeley PGA bioinformatics group for the creation of the CV gene database. This work was supported in part by the NIH-NHLBI Programmes for Genomic Applications Grant HL66728 and NIH Grant HL071954A through the US Department of Energy under contract no. DE-AC03-76SF00098.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Waterston RH, Lindblad-Toh K, Birney E *et al.*: Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-562 (2002).

- • **Landmark paper on the draft sequence of the mouse genome and its crucial role in understanding human biology.**

2. Aparicio S, Chapman J, Stupka E *et al.*: Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 297, 1301-1310 (2002).
3. Dermitzakis ET, Reymond A, Lyle R *et al.*: Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* 420, 578-582 (2002).

- • **Computational and experimental analyses of a large number of human–mouse conserved sequences with unknown function indicate that most of the sequences are non-genic and are suggestive of being involved in regulatory or structural functions.**

4. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC: Cross-species sequence comparisons, a review of methods and available resources. *Genome Res.* 13, 1-12 (2003).
5. Ureta-Vidal A, Ettwiller L, Birney E: Comparative genomics, genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* 4, 251-262 (2003).
6. Pennacchio LA, Rubin EM: Comparative genomic tools and databases, providing insights into the human genome. *J. Clin. Invest.* 111, 1099-1106 (2003).
7. Thomas JW, Prasad AB, Summers TJ *et al.*: Parallel construction of orthologous sequence-ready clone contig maps in multiple species. *Genome Res.* 12, 1277-1285 (2002).
8. Dehal P, Predki P, Olsen AS *et al.*: Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* 293, 104-111 (2001)
9. Ross MT, LaBrie S, McPherson J, Stanton VP Jr: Screening large-insert libraries by hybridization. In: *Current Protocols in Human Genetics*. Boyl A (Ed.), 5.6.1-5.6.52 (1999).
10. Murphy WJ, Stanyon R, O'Brien SJ: Evolution of mammalian genome organization inferred from comparative gene mapping. *Genome Biol.* 2, REVIEWS0005 (2001).
11. Groenen MA, Cheng HH, Bumstead N *et al.*: A consensus linkage map of the chicken genome. *Genome Res.* 10, 137-147 (2000).
12. Schwartz S, Zhang Z, Frazer KA *et al.*: PipMaker-a web server for aligning two genomic DNA sequences. *Genome Res.* 10, 577-586 (2000).
13. Gottgens B, Barton LM, Gilbert JG *et al.*: Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat. Biotechnol.* 18, 181-186. Erratum in, *Nat. Biotechnol.* 18, 1021 (2000).
14. Jareborg N, Durbin R: Alfresco – a workbench for comparative genomic sequence analysis. *Genome Res.* 10, 1148-1157 (2000).
15. Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES: Human and mouse gene structure, comparative analysis and application to exon prediction. *Genome Res.* 10, 950-958 (2000).
16. Bray N, Dubchak I, Pachter L: AVID, A Global Alignment Program. *Genome Res.* 13, 97-102 (2003).
17. Brudno M, Do CB, Cooper GM *et al.* and NISC Comparative Sequencing Program: LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 13, 721-31 (2003).
18. Brudno M, Malde S, Poliakov A *et al.*: Glocal alignment: finding rearrangements during alignment. *Bioinformatics* 19(Suppl. 1), I54-I62 (2003).
19. Mayor C, Brudno M, Schwartz JR *et al.*: VISTA, Visualizing Global DNA Sequence Alignments of Arbitrary Length. *Bioinformatics* 16,1046-1047 (2000).
20. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM: rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* 12, 832-839 (2002).
21. Xuan Z, Wang J, Zhang MQ: Computational comparison of two mouse draft genomes and the human golden path. *Genome Biol.* 4,R1 (2003).
22. Couronne O, Poliakov A, Bray N *et al.*: Strategies and tools for whole-genome alignments. *Genome Res.* 13, 73-80 (2003).
23. Oeltjen JC, Malley TM, Muzny DM, Miller W, Gibbs RA, Belmont JW: Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.* 7, 315-329 (1997).
24. Kuo CL, Chen ML, Wang K *et al.*: A conserved sequence block in murine and human T cell receptor (TCR) alpha region is a composite element that enhances TCR alpha enhancer activity and binds multiple nuclear factors. *Proc. Natl. Acad. Sci. USA*. 95, 3839-3844 (1998).
25. Loots GG, Locksley RM, Blankespoor CM *et al.*: Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288, 136-140 (2000).
26. Wolfe KH, Sharp PM, Li WH: Mutation rates differ among regions of the mammalian genome. *Nature* 337, 283-285 (1989).

- •• **Study that shows the rate of silent substitution varies among genes and is correlated with the base composition of genes and their flanking DNA.**

27. Smith JR, Osborne TF, Goldstein JL, Brown MS: Identification of nucleotides responsible for enhancer activity of sterol regulatory element in low density lipoprotein receptor gene. *J. Biol. Chem.* 265, 2306-2310 (1990).
28. Sekar N, Veldhuis JD: Concerted transcriptional activation of the low density lipoprotein receptor gene by insulin and luteinizing hormone in cultured porcine granulosa-luteal cells, possible convergence of protein kinase a, phosphatidylinositol 3-kinase, and mitogen-activated protein kinase signaling pathways. *Endocrinology* 142, 2921-2928 (2001).
29. Goodman M: The genomic record of humankind's evolutionary roots. *Am. J. Hum. Genet.* 64, 31-39 (1999).
30. Lin Q, Schwarz J, Bucana C, Olson EN: Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C. *Science* 276, 1404-1407 (1997).
31. Repa JJ, Mangelsdorf DJ: The role of orphan nuclear receptors in the regulation of cholesterol homeostasis. *Ann. Rev. Cell Dev. Biol.* 16, 459-481 (2000).
32. Whitney KD, Watson MA, Goodwin B *et al.*: Liver X receptor (LXR) regulation of the LXRalpha gene in human macrophages. *J. Biol. Chem.* 276, 43509-43515 (2001).
33. Modrek B, Lee CJ: Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* 34, 177-180 (2003).

- •• **A survey of the human, mouse and rat exons shows that most of the alternative spliced forms of exons in these genomes are species-specific and will not be readily detected by sequence comparison.**

34. Boffelli D, McAuliffe J, Ovcharenko D *et al.*: Phylogenetic shadowing of primate sequences to find functional regions of the

human genome. *Science* 299, 1391-1394 (2003).

•• **The first paper demonstrating that functionally conserved sequences can be detected by multiple closely-related species comparisons.**

Websites

101. http://www.americanheart.org/presenter.jhtml?identifier=3000090
American Heart Association. Heart Disease and Stroke Statistics – 2003 Update.
102. http://www.sciencemag.org/cgi/reprint/300/5618/409.pdf
Science article.
103. http://hgsc.bcm.tmc.edu/projects/rat/
BCM Rat Genome Project.
104. http://www.sanger.ac.uk/Projects/D_rerio/
Sanger Zebrafish Genome Project.
105. http://pga.lbl.gov/cvcgd.html
CV Comparative Genomic Database.
106. http://www.genome.gov/10001852
NHGRI BAC libraries.
107. http://www.nsf.gov/bio/pubs/awards/bachome.htm
NSF BAC libraries.
108. http://pipeline.lbl.gov/
VISTA Genome Browser.
109. http://genome.ucsc.edu/
UCSC Genome Browser.
110. http://bio.cse.psu.edu/pipmaker/
PIPMAKER.